Reflections on coding 90 million historical occupations

Evan Roberts
Minnesota Population Center, University of Minnesota

31st Social Science History Association Conference
Minneapolis, MN
2-5 November 2006

Abstract: The North Atlantic Population Project--bringing together late nineteenth century census records of 90 million individuals from North America, Britain, Norway and Iceland, faced the daunting task of consistently classifying more than 2 million distinct occupations in four different languages, with regional variation in occupational terminology in English, the most common language of respondents. This paper provides a retrospective on the task, and provides some generalizations from our experience that other researchers working with historical occupations may find useful. We coded our occupations into a modified version of HISCO. Our modifications to HISCO were primarily designed to reduce the number of codes in use, simplify the classification of laborers and other semi- and un-skilled workers who did not provide much information on the specific tasks they were engaged in, and provide more definite headings for commonly used occupational responses that HISCO provided more flexible codes.

Our coding project was successful. We achieved consistent classification across all five countries, and coded all occupations in the dataset into the modified HISCO scheme. HISCO's addition of subsidiary variables (STATUS, RELATE, PRODUCT) was a useful method for retaining information in occupational responses that would otherwise have been lost. The hierarchical structure of the codes simplified coding and consistency checking. HISCO would be improved by adding codes for industry that would allow separation of occupations that occur in different industries but are otherwise reported as the same job. Our use of the contemporary UN product classification for 19th century occupations was surprisingly successful.

Introduction

The North Atlantic Population Project (NAPP) is a collaboration of researchers at eight institutions in six countries to create a harmonized series of census microdata from late-nineteenth century censuses. Currently the data available contains the records of 89 million people from eight late-nineteenth century censuses. When complete, there will be records of another 28 million individuals. One of the most important and challenging set of variables to code have been the census questions related to work and employment, particularly occupational classification. Because NAPP has been creating public-use datasets we placed a high priority on using an occupational classification scheme that was well known, would be accepted by other scholars, and easily integrated with other data sources. To this end, we chose to adapt the Historical International Standard Classification of Occupations (HISCO) scheme that has itself been adapted from the 1968 version of the International Standard Classification of Occupations (ISCO, unsurprisingly). In 2003, at the beginning of our occupational classification project, we published an article in *Historical Methods* that described our modifications to HISCO and our intentions for efficiently coding more than two million occupational strings.[1] This paper provides a retrospective on what we have achieved in the past four years.

Background: The North Atlantic Population Project

The North Atlantic Population Project is a harmonized dataset of census microdata from six countries on the North Atlantic rim. It currently contains 89 million records from five complete count censuses from four countries, and two census samples—nearly all from the late nineteenth century (specifically, 1865-1901). When complete, it will contain the records of 127 million individuals from six countries, spanning the years 1703 to 1930 (see Table 1).

---

[1] Our 2003 article was jointly authored by Lisa Dillon, Chad Ronnander, Matthew Woollard and Gunnar Thorvaldsen and myself. While I have benefited tremendously from their counsel over the past four years, and their experience has made me wiser and better in this field, the traditional caveats about single authored papers apply here. Their views have improved this paper, but they are not to blame for anything inaccurate or controversial contained herein.

Table 1. Phase I and Phase II NAPP datasets

Census Year	Country	Sample Density	Number of Cases (thousands) Household	Person
Existing NAPP censuses (NAPP Phase I)				
1881	Great Britain	1.00	6,188	29,866
1881	Canada	1.00	799	4,278
1870	Iceland	1.00	11	60
1880	Iceland	1.00	14	72
1901	Iceland	1.00	15	78
1865	Norway	1.00	387	1,702
1900	Norway	1.00	395	2,294
1880	United States	1.00	10,138	50,486
TOTAL EXISTING			17,933	88,764
Censuses to be added (NAPP Phase II)				
1851	Britain	0.02	83	398
1852	Canada	0.05	31	170
1871	Canada	0.01	13	62
1891	Canada	0.05	67	350
1901	Canada	0.05	51	265
1911	Canada	0.05	74	372
1921	Canada	0.04	74	362
1931	Canada	0.03	67	320
1941	Canada	0.03	77	355
1951	Canada	0.03	93	420
1703	Iceland	1.00	9	50
1835	Iceland	1.00	10	56
1845	Iceland	1.00	10	57
1801	Norway	1.00	164	879
1875	Norway*	0.02	135	639
1890	Sweden	1.00	965	4,576
1850	United States	0.01	37	198
1860	United States	0.01	66	354
1870	United States	0.01	80	428
1880	United States	0.10	1,014	5,049
1900	United States	0.06	1,248	5,220
1910	United States	0.01	311	1,271
1920	United States	0.01	257	1,037
1930	United States	0.06	1,670	6,160
TOTAL TO BE ADDED			6,632	29,120

Note: Shaded cells represent census data currently available.

NAPP is a collaboration between researchers at institutions in each country involved in the project. The Minnesota Population Center at the University of Minnesota co-ordinates the project, disseminates the data, and does much of the final programming required to harmonize and distribute the data. However, the project is a true collaboration and without the input and efforts of all the individuals and institutions involved, making this one of the rare cases where the cliché about the whole being greater than the sum of its parts quite true. Specifically, the principal collaborators involved in the project are listed in Table 2.

Table 2. NAPP participants

Country	Institution	Principal collaborators	Responsibility (Year indicates a census)
Canada	Ottawa	Chad Gaffield	1911, 1921, 1931, 1941, 1951
	Montreal	Lisa Dillon	1852, (1861[2]) 1881
	Guelph	Kris Inwood	1871, 1891
	York	Gordon Darroch	1871
	Victoria	Peter Baskerville	1901
Great Britain	Essex	Kevin Schürer Matthew Woollard	1881
Iceland	Statistics Iceland	Ólöf Garðarsðottir	1703, 1835, 1845, 1870, 1880, 1900
Norway	Bergen	Jan Oldervoll	1801, 1865
	Tromsø	Gunnar Thorvaldsen Marianne Jarnæs-Erikstad	1875, 1900
Sweden	Umea	Per Axelsson Mats Danielsson	1890, 1900
United States	Minnesota	Steven Ruggles (PI) Ron Goeken Evan Roberts (coordinator) Sula Sarkar (research assistant)	1850, 1860, 1870, 1880, 1900, 1910, 1920, 1930
	Northwestern	Joseph Ferrie	Data linking

Content and comparability of nineteenth century censuses

Integrating nineteenth-century census data from the NAPP countries into one dataset turned out to be a relatively simple and straightforward task in concept and execution. The censuses were taken in largely similar ways, because of extensive

[2] Tentative future project.

correspondence among contemporary census officials in the different countries.[3] The questions asked in censuses were also very similar.

The most outstanding difference among the censuses *across* countries was that the British census of 1881 was a *de facto* census, and the Swedish census of 1890 was compiled not by a door-to-door enumeration, but by capturing parish registration records for a point-in-time into a census. The other censuses were entirely *de jure* enumerations, or a mixture of both *de facto* and *de jure* enumeration.[4] On an *a priori* basis the difference in enumeration methods should have little impact on the enumeration of occupations. Indeed, one might expect the British *de jure* enumeration to be somewhat more accurate for transient occupations since the census taker was more likely to encounter the individual themselves.

The chances of the individual speaking to the enumerator or filling out the census form themselves—not the *de facto* or *de jure* enumeration rule—was the most significant influence on the accuracy of the information collected in the nineteenth century censuses. This also varied between the different NAPP countries. Compared to present day censuses, the enumeration procedures in some of the late nineteenth century censuses meant that enumerators were less likely to get accurate information from all individuals, because not all individuals filled out their own census form, and the enumerator did not speak to every individual they enumerated. Specifically, in Canada and the United States the enumeration was conducted by census enumerators—often local officials, and in the United States particularly likely to be connected with current local *elected* office holders—who visited households over the course of several weeks.[5] While enumerators would return to households that were empty upon a first visit, they did not seek to speak to all individuals. The consequences for the accuracy of the census were that individuals with weak social ties to the primary family in the household—boarders and lodgers, for example—and individuals not present when the enumerator visited were likely to have

---

[3] Add note from original article.

[4] The difference between *de facto* and *de jure* enumeration is as follows. A *de facto* enumeration counts people in the place they are found on census day, whether that is their permanent residence or not. A *de jure* enumeration counts people at their permanent residence. Thorvaldsen, *Historical Methods*, spring 2006.

[5] Further information on the American census procedures is available in Diana Magnuson and Miriam King, "Enumeration Procedures," *Historical Methods* 28, no. 1 (1995): 27-32.

information recorded imprecisely. In the case of precisely specified variables such as age, the information provided about absent household members can be said to be wrong, at least if we make the conventional demand that ages be given to the nearest whole year, rather than nearest five years. The consequences for the age variable, for example, is a marked degree of age heaping in the 1880 American census, beyond what can be attributed to our supposition that nineteenth century society was somewhat less numerate than contemporary society.[6]

Ages, even if they were sometimes wrong, at least conform to a comparatively small range of values. Most of the other variables collected by the nineteenth century censuses had a naturally small range of possible values, and thus required little explicit instruction that respondents limit the range of their responses. Sex and age are the best and most common examples.

Alternatively, where the number of possible responses was potentially large the instructions to enumerators directed the collection of information in a way that limited the number of possible responses. Birthplace is a good example of this type of variable. Were respondents allowed to give an open-ended answer, the number of possible responses would swell to at least the combined number of towns, villages and other small administrative units, but would probably also include colloquial references and even more specific data that people were born on particular farms or particular ships at sea. The flipside of this phenomena—people who do not know where they were born—does not create many more unique responses to the question, since there are a limited number of ways to say "I don't know". Seeking to avoid this proliferation of responses, census officials in all the NAPP countries directed that enumerators collect a sub-national geographic area as the birthplace for the native born (a province or parish or municipality), and a country for the foreign born. This limits the number of responses to a more manageable several hundred, except in Great Britain where there were 15,000 odd parishes. Unsurprisingly, many people did not know their parish of birth and gave as responses the names of small towns and cities that they thought to be parishes, but were

---

[6] In general, see Patricia Cline Cohen, *A calculating people: the spread of numeracy in early America* (Chicago: University of Chicago Press, 1982). and Michael A. Bernstein, "Numerable Knowledge and its Discontents." *Reviews in American History* 18, no. 2 (1990): 151-164.

not. Yet even this type of variation is relatively straightforward to code and classify, though I do not understate the time involved.

The process of taking the data from the mouths of respondents to electronically coded datasets introduces several points where errors can be introduced, particularly at the transcription stage. Many of the unique and different responses that we observe in the NAPP census databases are respondent uncertainty and transcription errors, but the vast majority of people provide clear answers to census questions that conform to the limited range of values requested by the census agencies.

Occupation questions, conversely, had a much wider range of potential values and deliberately so. While it is likely the delusion of every age to believe that it is living in tremendously changing times, few now would disagree that the late nineteenth century was a time of significant social and economic change in the countries on the North Atlantic rim. This perception of change in production methods and industrial structure was noted by contemporaries, and contributed to census agencies motivation to record in great detail the occupations and trades of their peoples.[7] The American instructions to enumerators are worth quoting in full:

> The inquiry "profession, occupation, or trade," is one of the most important questions of the schedule. Make a study of it. Take especial pains to avoid unmeaning terms, or such as are too general to convey a definite idea of the occupation. Call no man a "factory hand," or a "mill operative." State the kind of a mill or factory. The better form of expression would be, "Works in a cotton mill," "Works in paper mill," etc. Do not call a man a "shoemaker," "bootmaker," unless he makes the entire boot or shoe in a small shop. If he works in (or for) a boot or shoe factory, say so.[8]

Similarly, in Great Britain the instructions to enumerators requested that they collected information on the physical size of farms, and the numbers employed by farmers and manufacturers. The consequence was a rich collection of information about the occupations and industries of 90 million people. In Great Britain there were more than 1.3 million unique and different responses to the occupation questions in the census,

---

[7] See e.g.; R.T. Ely, *An Introduction to Political Economy* (New York: Chautaqua Press, 1889): 491.
[8] See http://usa.ipums.org/usa/voliii/inst1880.shtml.

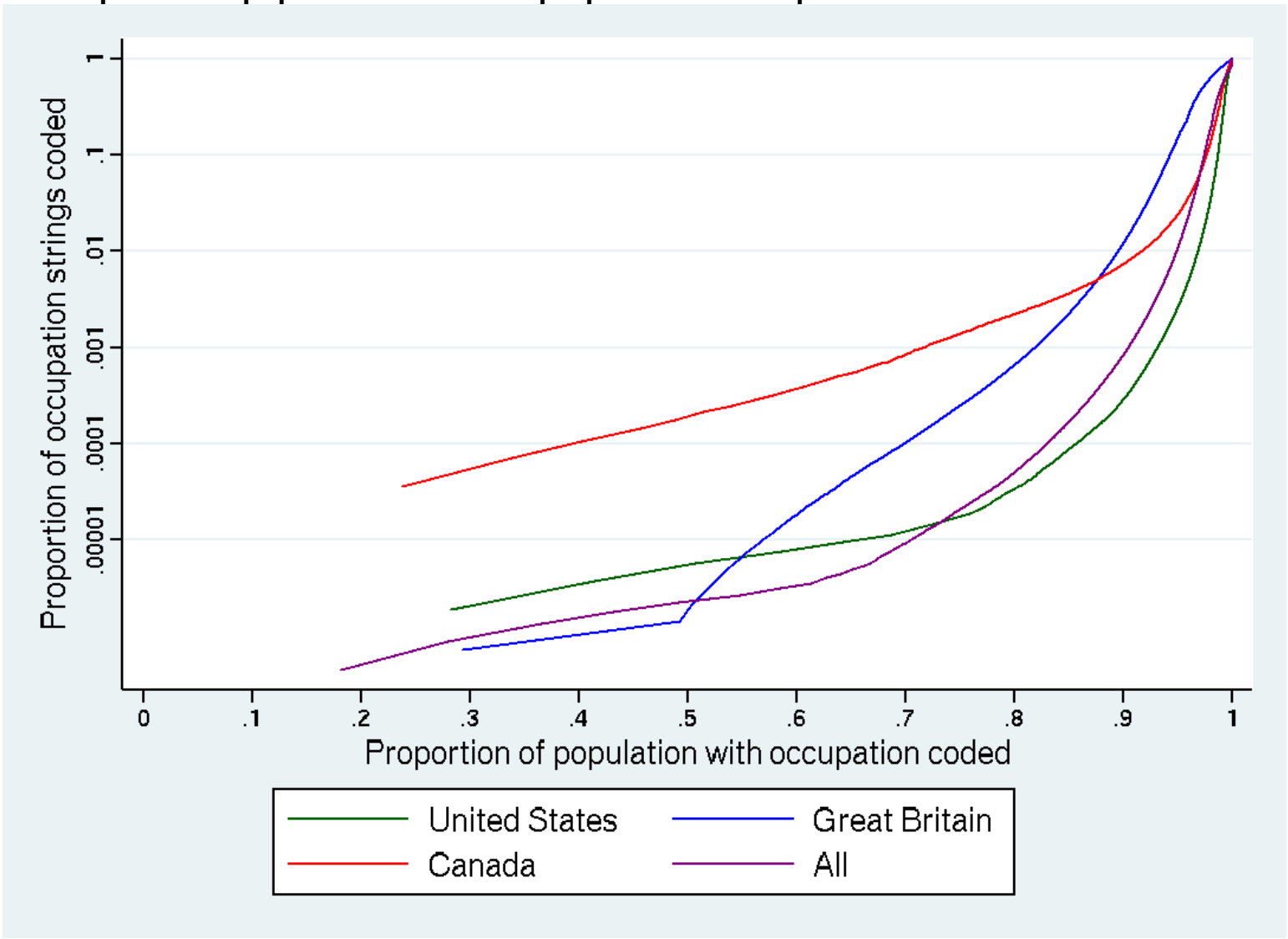
more than 500,000 in the United States, 367,000 in the two Norwegian censuses, and 28,000 in the Canadian censuses. In total we had more than two million occupations to code and classify—a task now largely completed. As we add more samples to the NAPP database, we will have hundreds of thousands of additional occupations to code, particularly from the twentieth century American and Canadian censuses, and any future samples of nineteenth century British censuses we can acquire.

The number of distinct occupational responses (also referred to as "strings" or "unique responses") gives an accurate indication of the scale of our task in coding and classifying *all* the occupations. However, it gives a misleading impression of quite how much diversity there is in occupational structure in any of the countries. The vast majority of individuals receive a valid occupation code when a small fraction of the different responses have received codes, because some responses are very common. For example, 4 million people give the same simple response "Farmer" in the United States, along with 400,000 in Canada. This represents one in every nine people in the United States who had an occupation, and nearly one in four (23.8 per cent) of the people giving an occupation in Canada (See Figure 1).[9] With just one per cent of the occupation strings, we have assigned codes to more than 85 per cent of the population in all countries (both separately and combined).

All of the most common responses to the occupation question are straightforward to assign occupational codes to. Thus, we made very rapid progress initially, and spent the remaining years of the project coding the occupational responses of approximately two million people whose occupational responses were infrequent, and not shared with others. The cliché that "the devil is in the details" truly applied to our work here, with the devil also lurking in the tails of the distribution, where every string assigned a code represented the response of one person only. In the United States—which I can speak about more authoritatively since I personally coded about 400,000 of the 575,000 strings in our dictionary[10]—a substantial minority of the strings with frequency 1 were gibberish,

[9] My thanks to Matthew Woollard for producing the first version of this figure, which I have re-created for this presentation.
[10] The remainder were coded by Chad Ronnander.

Figure 1. Proportion of population coded for proportion of occupations coded



so mangled in the process of getting from the respondents mouth to the census enumeration sheets, to microfilm, to the LDS transcription, and finally to our electronic version of the data that we were unable to assign any code except "Occupation unknown" (code 99500). In the course of coding the occupations for the whole population this, of course, took some time, since we had to look at each string, decide that it was gibberish, assign a code, and move on. Multiply several seconds by 40,000 and you have a good full-time week of looking at poorly transcribed occupations. Yet the conceptual challenge of a lot of gibberish strings was not great, and procedures for handling this kind of data are well-established. The unsurprising lesson of having the opportunity to code all the occupations for five complete-count censuses is that you see a lot of them. To generalize for future researchers, the tradeoff of collecting a large sample—all the way up to a whole population—is that you can spend less time on each case. The gibberish we saw could well have been introduced at a point where it would still be possible to correct it by looking again at the microfilm. But in the context of coding millions of different occupations on a relatively small budget, we felt that having to assign a code for "Unknown or ambiguous occupation" to less than 1/10 of one per cent of the population was a very acceptable tradeoff.[11] At the outset of the project we anticipated that the scale of the coding task in front of us might require that we assign ambiguous codes to many more occupations. Having been able to actually code the entire population is something of an accomplishment in itself.

Classification scheme

The size of our coding project influenced our choice of classification scheme. Although many people use "coding" and "classification" as loosely synonymous, we prefer to think of "classification" as the global process of deciding what dimensions of occupations are important and how those occupations will be grouped and separated. Coding is the application of a classification scheme to individual occupational responses. We had a lot to code, and while it could be semi-automated, the choice of classification scheme would influence how quickly we could assign codes.

[11] In the United States 43,929 people out of 50,491,088 received code 99500 for unknown occupation.

At the time we began collaborating on NAPP occupational codes had been assigned to all of the British and Norwegian data, and about one-third of the U.S. data.[12] Only the small number of Canadian and Icelandic occupations had not been coded. In some ways this pre-coding into different schemes was a disadvantage, since we would be repeating work we had already done. Moreover, Chad Ronnander and I felt quite strongly that the United States 1950 classification scheme which we had extensive experience in coding with, was easily applicable to historical data with some well known and limited exceptions (principally relating to teamsters working with horses, the structural equivalent of modern drivers, even if their specific tasks differed significantly). Moreover, although the United States 1950 scheme was designed by the United States Census Bureau and has some American peculiarities, it could easily be applied to data from other countries (and is in fact being used by the Canadian Century Research Infrastructure for their samples of Canadian censuses from 1911 to 1951 that will be available through NAPP in the future). Using a well-known classification scheme was an important consideration, and by virtue of its existing use in the IPUMS the United States 1950 scheme was well known. The domestic Norwegian and British schemes were less well known, and less easily applied to other countries. Moreover, the British classification scheme was the original one from 1881 which made some decisions peculiar to modern minds. Grave diggers and clergy men, for example, received the same code. While this decision may seem peculiar, confusing occupation and industry, this kind of conflict over which aspects of a job should receive the most priority in assigning codes, is hardly unique to the Registrar General's Office in 1881. When any combination of tasks and roles described in a word or phrase is given one code these kinds of conflicts and trade offs are inevitable.

In both Norway and the United States where the domestic coding scheme incorporates both an occupational code, and an industry or trade code, many though not all, of these conflicts are reduced. Adding an industry code to our classification scheme would have increased the scale of our task by approximately one-third, though in the United States and Norway we would have been able to convert many industry codes by programming. In an ideal world we would have added industry codes. This is perhaps the

[12] All of these domestic codes are available in the final release of the NAPP data.

most substantial lesson that I have learned, or had confirmed, during our coding project: coding both occupation and industry should be a priority. We will return to this point throughout the paper. Not coding industry was a choice assisted by three factors:

1. Amount of time available for coding
2. Industry was not specifically enumerated in any census
3. The classification scheme we used did not explicitly incorporate industrial codes

As the foregoing indicates we considered the following factors important when choosing a classification scheme

- Ease of application in coding
- Acceptance in the scholarly community
- International comparability

Having rejected application of any of the domestic coding schemes to other countries in NAPP, we briefly considered using a modern international coding scheme such as ISCO, but were soon drawn to the HISCO scheme then under development by an international group of scholars.[13] HISCO was a modification of ISCO—the International Standard Classification of Occupations—and thus easily met two of our three main criteria: acceptance in the scholarly community and international comparability. Thus, our main considerations in evaluating HISCO was how easily we could apply it in coding two million occupations from multiple countries by several different coders. We were concerned that within an individual country—particularly Canada and the United States where several staff would be coding—and across the different countries, that we would give the same distinct occupational response or string the exact same code.

Because so many of the individuals in each country received a code from coding a limited number of strings we repeatedly compared our coding of a small combined list of occupations. For the English speaking countries (Canada, Great Britain and the United States) this list contained any occupation that was in the top 1000 most frequent occupations in any country individually, and any occupations in the top 1000 most

[13] Marco H.D. van Leeuwen, Ineke Maas, and Andrew Miles, "Creating a Historical International Standard Classification of Occupations," *Historical Methods* 37, no. 4 (2004): 186-197, Marco H.D. van Leeuwen, Ineke Maas, and Andrew Miles, *Historical International Standard Classification of Occupations* (Leuven: Leuven University Press, 2002).

frequent occupations for all countries combined (See Appendix 1 for the top 100 version of this list).[14] Through this process we discovered that while HISCO was easy to apply for most occupations it was not easy to consistently apply it to the following broad categories of occupations

- Managers (Major groups 2 and 5)
- Laborers and unskilled operatives in factories (particularly in the United States) (Major groups 7-9)
- Vague or industry only responses (e.g; "Works on railroad")

Lest this sound entirely like criticism it must be said that for several groups, covering a large proportion of the occupation, HISCO proved eminently usable. In these groups our main modifications were in the spirit of what ISCO and HISCO propose, which is to take advantage of the hierarchical structure of the coding scheme. These groups were

- Professionals (Major groups 0 and 1)
- Clerical Workers (Major group 3)
- Sales workers (Major group 4)
- Agricultural workers (Major group 6)

Our changes in line with the spirit of HISCO—aggregating up the hierarchical structure of the codes—were primarily in major groups 0,1, 3, 4 and 6. Especially with professionals the precise distinctions made in HISCO between, for example, organic and inorganic chemists, never appeared in the data. Rather than implying that they did by using this level of precision we preferred to aggregate these types of responses into one group.

In the clerical and agricultural groups we retained much of the structure of HISCO, but introduced new headings (or 5 digit codes) to reflect very common responses that often introduced industrial distinctions between occupations. For example, we distinguished b between timekeepers in factories, on railroads and elsewhere. In a slightly different way,

[14] This list is available on request, as is the whole occupational dictionary.

in the agricultural group we added codes for particular, perhaps peculiar, nationally specific responses that we wanted to preserve, such as "habitant" and "farmer and fisherman." In part these decisions reflected a lack of confidence that we really knew what some occupations were about.

This brings me to another reflection or lesson from this process. Censuses are not well-placed to tell us much about "what do people do all day."[15] Occupations are a set of functions, tasks and social roles that people perform within the context of what we might generically call "firms." Sometimes (perhaps often in the nineteenth century) those firms overlapped with families and farms, but it pays to think of "firms" as distinct social groups. It is quite clear from the research of business and labor historians that people within the same firms, with the same job title, may nevertheless be doing quite (I use this word in its British sense, though aware of its ambiguity) different tasks, and have quite different social relations to their colleagues and employers. Nevertheless people then take these occupational titles into the wider world, into other social contexts, as part of their social identity. This much is also clear from social-historical and contemporary research: the work people do is a significant part of their self-image and the image others have of them, which contributes to their social status in the community. Yet all we have for most people in the census (literally most, see Appendix 1) is a one word description of their occupation. Farmer, lawyer, teacher, blacksmith, for example.

In some cases, the lack of detail about what people did was a reflection of the enumeration process. Farmers did not provide extra detail on their work in Canada and the United States because there was a simultaneously conducted agricultural census that collected this information, and similarly for manufacturing employers. The potential to link these sources to NAPP in future is great, and will considerably expand our knowledge of the connections between population and economy, but for now we really have only the briefest of descriptions about most people's occupations. While the majority of strings provide more detail, the vast majority of people do not.

Being able to see the entire range of responses for these countries put us in a somewhat unique position, and influenced our thinking about the way in which we can interpret occupational codes. Loosely, we refer to our approach as "nominalist." All we

[15] Some of you may recognize this phrase as due to the children's author Richard Scarry.

think we can do is group relatively similar nominal phrases. Providing extra interpretation is the job of researchers who may be able to link individuals to other sources—business records, for example—that provide more detail on what people were doing.

What we call a "nominal" approach often coincides with a "functional" approach, where an occupation is defined by what a worker does for other people within a system of production, not by task or hierarchy. If one is interested in creating an occupational classification system that allows one to track long-term change, then a functional or nominalist approach avoids frequent discontinuities in the scheme. Technology changes rapidly enough that if we try to pin down exactly what tasks an occupation involves we will not be able to create a truly historical classification scheme. Some occupations will emerge and decline, and introduce new titles to the lexicon of occupations, even with a nominal/functional approach. While tasks have changed tremendously in the last *x* years—I am being purposefully vague here—the number of truly new occupations in the last 150 years is tiny.

This approach guided our modifications to the "vague" titles and the primarily industrial responses that we particularly observed in the United States. Responses such as "works in cotton mill" or "works on railroad" were particularly common. We made the assumption that these workers were production workers, rather than clerical workers, because clerical workers were relatively rare, and would, we guess, have been more likely to view themselves as clerical workers, not workers in a cotton mill who happened to be primarily writing and filing. These assumptions are also backed up by what we know of the structure of manufacturing and transportation firms from business records. Our approach to coding these responses was to create a nominal code for these individuals, close in the hierarchy to more specific production occupations in the same industry.

Our modifications reduced the number of codes we would apply to the data from 1,881 headings (5 digit codes) in HISCO to 650 codes in our modified version. We left space in the scheme to add in new codes during the coding process if we discovered high-frequency occupations that deserved their own code. In practice, this did not occur very often. We created codes to distinguish between different types of ticket and baggage

agents in transportation, and clarified our codes for workers on railroads. Otherwise, the modifications we made before beginning the majority of the coding stood the test of applying them to millions of records.

Practically coding two million occupations

The practical details of coding this many occupations are not exciting, but a brief recapitulation of how we did that successfully and quickly might be useful. As mentioned, 90 million people gave 2 million distinct responses to the occupational question. We integrated these in an Access database, so that the duplication of common responses such as "Farmer" in the United States, Canada, and Great Britain did not result in three entries for identical occupation. While "only" 6,601 occupations overlapped in the three English-speaking countries these occupations covered a substantial proportion of the population. These overlapping occupations covered 35.8 million people in total, 23.8 million in the United States, 10.9 million in Britain, and 1.1 million in Canada. Indeed, the remaining 22,000 Canadian responses covered only 263,334 people—all the remaining people with an enumerated occupation. The British proportion covered by the common list is low because farmers and manufacturers returned their numbers of employees, and size of farms introducing more variation into the distinct responses. We refer to the Access database of unique occupational responses as a "dictionary," and I use this terminology in the rest of the paper.

The dictionary contained responses unique on only one variable (occupation), and when we were coding we saw no other information about the respondent. It would have been possible to create a dictionary that, for example, contained the unique combinations of occupational response and the name or type of institution (if any) the individual was living in. As most individuals did not live in an institution this would not have added many entries to the dictionary, but may have provided valuable extra information on interpreting ambiguous occupations. A "keeper," for example would be assigned a different occupation code if we could see that he was living in a jail. Without this information we assume a keeper is a retail store keeper. If we were to do the project over, incorporating group quarters information in the dictionary would be something to

consider, though we plan to identify some inconsistencies through programming given the size of the dataset. Similarly, it would be useful to know whether women who responded that they were "keeping house" were living with their family or not. This type of information could have been included in the dictionary requiring coders to make a decision about occupational codes conditional on other responses. Again, the size of the dataset makes it more practical to fix these kinds of problems with post-hoc programming.

Researchers embarking on occupational coding projects with later censuses or surveys often have separate information on occupation and industry, and the dictionary would then be created on the unique *combinations* of occupation and industry. In some cases the correct occupational code is dependent on industry. Although the nineteenth century censuses asked just one question which purported to be about occupation, in practice the questions elicited both occupational and industrial information from most of the population. In the United States just nine per cent of the 1880 population with an occupation could not also be assigned an industry code (though "just" nine per cent of more than 30 million workers is still 2.8 million workers without an industry).

The bottom line is that the structure of our dictionary was very simple: it contained unique occupations. Although we were creating a harmonized dataset, we created country-specific versions of every variable in the HISCO classification scheme. As well as the primary variable for occupation, HISCO also includes variables for additional information in the occupation string about social status (OCSTATUS in NAPP) and relationship to someone else's occupation (e.g, the common response "Farmer's wife" or "Farmer's son." OCRELATE in NAPP), and the product sold by retail workers. As well as fields for entering codes, the dictionary also included the following variables

- Frequency of the response in every country
- A field for standardizing spelling and the format of the occupational response, so far only used in the United States. This information is available as OCCLABEL in NAPP.[16]

[16] While most entries are spelled correctly, some occupations are littered with spelling and transcription mistakes. Physicians and veterinarians are the best examples. While these occupations have their own code,

- Utility fields that we used to create copies of the occupational strings for editing and sorting. For example, sometimes the key word that is needed to assign a code does not appear at the start of the string. In a utility field we could modify the string to get the important words at the beginning of the string and then sort on this modified string to group similar occupations.
- Fields for marking who assigned codes. This was primarily used by the coding teams in Canada and the United States, where two or more people worked on the data simultaneously.
- Fields for flagging a response for follow up and discussion. As mentioned previously, we thought we might add more codes in the process of coding, which would have required some people to re-assign some codes. As it turned out, we did not have to revisit many coding decisions.

Copies of the integrated dictionary were distributed to all participating groups in late 2002, and we began coding in the integrated dictionary in early 2003. At the beginning of the project we proposed that the coding would be integrated in real time in an internet-based "collaboratory." While this lovely neologism may have assisted us in getting the project funded, in practice the development and bandwidth costs of the "collaboratory" were prohibitively expensive. We fell back on maintaining a central version of the data at Minnesota, and having the other national coding teams send us periodic updates that were merged into the central dictionary and then redistributed. Although these updates took several hours, and the data had to then be re-distributed to the participants, requiring that they stop coding for a day or two, this procedure proved to be effective and an economical use of our time. Much as everyone loves occupational coding, no one does it literally full-time.

In part this is because actually coding data is a task—an occupation, perhaps—difficult to do accurately for hours on end. The "mechanics" of coding took two forms. Working from the principle that we coded the first occupation unless a second occupation

---

some distinct occupations with spelling mistakes are subsumed into more general codes. Merchants and dealers, for example, receive the same code. "Merchant" was frequently abbreviated or spelled incorrectly in the American data.

modified it (e.g; farmer and legislator is coded as farmer, but broker and real estate agent is not coded as broker not specified), we coded a large number of strings using SQL update statements based on keywords appearing at the beginning of the occupation. Along with the small number of high frequency responses covering a large proportion of the population this meant that within a few weeks of starting to code we had assigned codes for well over 95% of the population. However this meant we had hundreds of thousands of occupations left to code by hand. We were able to expedite this procedure in Britain, Norway and the United States because of the existing domestic codes that had already been applied to the data. For some occupations there was a one-to-one or many-to-one relationship between the domestic codes and HISCO codes. The coding that took most of our time was for occupations that were

- Poorly transcribed or recorded, and thus could not be easily identified by keywords. We quickly were able to code common spelling mistakes or abbreviations as easily as if they were the correct words
- Given a domestic code that mapped to multiple HISCO codes. A good example of this were salespeople in the United States. In the 1950 occupational classification scheme we had already applied, salespeople in stores and traveling around largely receive the same code.[17] They are distinguished by their industry. Traveling salesmen are almost always *not* in retail stores. The HISCO codes were an improvement on the United States domestic codes in this respect since these are quite different occupations, and need to be distinguished.

This mix of automatic and manual coding proved both efficient and accurate for coding two million occupations. At the current time we have not yet incorporated the British occupations into our data as the coding there has not quite been completed.

Reflections and lessons from coding

Reflecting on this experience, the unique advantage of NAPP was that we were able to see the entire distribution of occupations for a whole country for one census. I

[17] Traveling salesmen in this period now have their first substantial monographic treatment. See Walter A. Friedman, *Birth of a salesman: the transformation of selling in America* (Cambridge, MA: Harvard University Press, 2004).

doubt that many people have been able to do this. Certainly the original coders of these censuses were not able to do so, since they lacked the easy ability to restrict themselves to the unique responses. As best we can tell from the publications of the Census Bureau, and examining the manuscript enumeration sheets the coding staff worked through the census line by line tallying the totals and assigning codes. There are many advantages to having the data on disk, and it really becomes quite feasible to peruse all the occupations.

Seeing all the occupations does not make my observations by themselves particularly insightful, but if you multiply the experience of seeing all the data by our limited insights, there are some things to say about enumeration and coding of occupations. One point, which may already be clear, is that instructions to enumerators matter. More generally, the original aims of the people collecting the data have a strong influence on the kind of responses that people provide. In a recent article, Vikström, Edvinsson and Brändström argue that:

> Any source containing occupational information reflects, in some way or the other, a situation of *negotiation* between the keeper of the register and the informant. This negotiating procedure involves two parties with different powers of influence. That power, or negotiating strength, will vary from situation to situation and from time to time. In one situation the informant can probably control the procedure and simply tell the record keeper what title to enter without having to support his or her claim with further evidence. In another situation the record keeper might enter a title without even having to confront the informant. The purpose of the registration should also be taken into consideration. In some situations a more precise occupation title is considered important while in other situations the occupation information is of less interest. Especially for some social classes the titles in the parish registers are often rather vague. They can be described as workers while we in other sources can find more precise titles as sawyer.[18]

When you see <u>all</u> the occupations the significance of different enumeration instructions and the context of census-taking becomes clearer. Some of the differences

---

[18] Par Vikström, Sorën Edvinsson, and Anders Brändström, "Longitudinal Databases — Sources For Analyzing The Life-Course: Characteristics, Difficulties And Possibilities," *History and Computing* 14, no. 1/2 (2006): 109-128.

between the types of responses that we observe, and the types of responses that the developers of HISCO saw, were because we were looking at census data exclusively. Particular instructions to enumerators in the United States and Great Britain revealed themselves in common responses. The United States instructions said that

> The organization of domestic service has not proceeded so far in this country as to render it worth while to make distinctions in the character of work. Report all as "domestic servants."[19]

Consequently we observe a great deal more detail in the enumeration of domestic service in Great Britain than in the United States. In Britain we are able to tell where domestic servants specialize in particular tasks such as cooking and cleaning. It seems highly unlikely that multiple-servant households in the United States did not have a similar specialization in tasks, but much of this is lost to history. In our modifications to the HISCO codes we annotated some codes with the note that a distinction between domestic and non-domestic cooks was only made in Britain, for example. Similarly, in Great Britain the instructions that enumerators record the physical size of farms, and the number of employees on farms and in manufacturing, was faithfully carried out by the enumerators. These forms of responses

> *Farmer of 317 Acres, employing 8 Labourers and 3 Boys*
>
> *Carpenter—Master, employing 6 men and 2 boys*

contain a great deal of useful information that we plan to parse and extract into several new variables. The syntax of the English language where responses always come in the form <number> <unit> make extracting the information a straightforward task. We expect that these variables, combined with detailed geographic information will be extremely valuable for the study of late-nineteenth century British economic history.[20]

Although this type of data is very valuable it is only available for British farmers and manufacturers. For most individuals we lack continuous measures of anything related to occupations that would differentiate between otherwise identical responses of

---

[19] http://usa.ipums.org/usa/voliii/inst1880.shtml

[20] James Foreman-Peck, *New perspectives on the late Victorian economy: essays in quantitative economic history, 1860-1914* (Cambridge: Cambridge University Press, 1991).

"agricultural labourer," "works in cotton mill," or "cultivateur" to give common examples from both sides of the Atlantic and both sides of the St. Lawrence River. Clearly in the census we have continuous measures of demographic characteristics—age, number of children, size of household, and can derive other measures from these—yet we lack any information on earnings, or hours worked.[21]

Thus, I am skeptical that we can do much besides code nominal occupational responses. Going much beyond this to impute socio-economic status or prestige for *individuals* is far more than *census data* can bear. While I am persuaded that there is impressive stability in the average status and earnings of occupations over time, that does not imply much about the variation between individuals at a point in time, which is often what we are interested in.[22]

Even within a single narrowly defined occupation, it is not uncommon to see annual earnings vary by a factor of two or three between the 5th and 95th percentiles of the earnings distribution.[23] We know that many of these earnings differences are due to the effects of age—which we do see in the census—but also to experience, tenure at the current workplace, and education. We lack measures of all of these variables in the census in the late nineteenth century. Implying ordinal or ratio properties to nominal occupational returns based on average earnings from other sources pushes the data a long way. With complete-count data, we have enough degrees of freedom that we can use occupational fixed-effects to control for the influence of specific occupations on other behavior measured by the census.

Similarly, I am skeptical that we could apply prestige measures to historical occupations. The literature suggests that the prestige of an occupation is negatively related to its frequency in the population, that is, the more people in a particular occupation the less prestige it has. But the very results of coding historical occupations show that particular occupations do not have a fixed place in the occupational structure

---

[21] We are adding the months unemployed variable to 10% of the U.S. population. This variable is available in the 1% sample.

[22] Matthew Sobek, "Work, Status, and Income: Men in the American Occupational Structure since the Late Nineteenth Century," *Social Science History* 20, no. 2 (1996): 169-207.

[23] Calculated from datasets in the Historical Labor Statistics Project. http://eh.net/databases/labor/. See Susan B. Carter, Roger L. Ransom, and Richard Sutch, "The Historical Labor Statistics Project at the University of California," *Historical Methods* 24, no. 1 (1991): 52-65.

over time, or across countries. While there may be stability over the course of a decade, or perhaps two, beyond that prestige of individual occupations shifts. Clerical work in the late nineteenth century was a relatively prestigious occupation because it required literacy, not yet a near universal skill, and in many firms clerical workers exercised real power in the running of the firm. Within 40 years, by the 1920s, clerical work was less prestigious as the skills required to perform clerical occupations became more widespread, and the tasks performed by clerical workers were sub-divided. The decline in status of "secretaries" from being executive officers of firms to handmaidens of middle management is the most dramatic consequence of this trend. Conversely, the prestige of nurses increased as the job required increasingly skilled training, and the term came to connote trained nurses in the medical field, rather than unskilled carers, similar to domestic servants. These are dramatic examples, but they illustrate the point.

As we expand NAPP to cover a much wider span of time, we will encounter many of the same occupations. The specific tasks done by people in these occupations will change in ways we do not know from the census. Different instructions to enumerators will induce important shifts in responses, that may appear to be changes in occupational structure. The best we can do is to code the responses we see accurately and consistently, and I hope that with NAPP we have done that.

Appendix 1. Top 100 occupations in the Canada, Great Britain and the United States

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
(Blank in the United States)			14,253,994	1			14,253,994	1
@@@ (British code for blank)					7,663,218	1	7,663,218	2
Keeping House			6,636,009	2	38	16832	6,636,047	3
Scholar	10,465	19	18,429	107	5,183,336	2	5,212,230	4
At Home	1	19673	4,674,864	3	8,826	178	4,683,691	5
Farmer	400,358	1	3,995,932	4	28,872	49	4,425,162	6
At School	5,480	33	2,873,713	5	2,083	619	2,881,276	7
Laborer	1	23972	2,225,742	6	42,151	35	2,267,894	8
Farm Laborer	2	7354	1,048,775	8	24,588	61	1,073,365	9
Keeps House			1,062,069	7	103	7515	1,062,172	10
Servant	43,179	7	742,678	10	52,704	24	838,561	11
Works On Farm	3	5213	793,328	9	208	4188	793,539	12
Carpenter	28,539	9	335,054	12	99,582	14	463,175	13
Farming	1,363	106	405,491	11	122	6573	406,976	14
Housekeeper	316	278	224,005	15	125,515	10	349,836	15
Coal Miner	28	1192	52,496	49	285,313	3	337,837	16
School	52,932	5	245,090	13	17,699	98	315,721	17
Labourer	91,827	4	24,714	89	197,341	6	313,882	18
Dressmaker	5,811	28	56,552	44	224,542	5	286,905	19
Ag Lab					284,033	4	284,033	20
House Keeper	1,337	107	219,100	16	26,544	51	246,981	21
Farm Hand	340	259	237,804	14	49	13689	238,193	22
Domestic Servant	197	362	88,529	29	147,430	8	236,156	23
Blacksmith	13,820	17	150,960	21	67,046	21	231,826	24
Keep House			212,767	17	19	30395	212,786	25
Home	3	6062	196,579	18	1,599	777	198,181	26
House Keeping	4	4618	168,373	19	75	9661	168,452	27

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
* (Canadian code for blank)	168,011	2	221	3020			168,232	28
Laundress	481	205	28,112	83	139,064	9	167,657	29
Cotton Weaver	41	931	1,665	628	165,502	7	167,208	30
Housekeeping			163,132	20	314	2991	163,446	31
Painter	5,772	29	91,405	27	45,511	33	142,688	32
Tailor	5,227	34	64,446	41	72,778	18	142,451	33
Dress Maker	5,515	32	83,943	31	49,847	29	139,305	34
Attending School	485	203	134,762	23	164	5126	135,411	35
Clerk In Store	102	530	135,069	22	4	120141	135,175	36
Cook	1,129	120	92,033	26	40,934	37	134,096	37
Miner	5,636	30	120,554	24	2,960	471	129,150	38
Clerk	18,363	12	70,636	36	38,930	41	127,929	39
Butcher	4,584	40	71,039	35	51,768	25	127,391	40
Annuitant					124,308	11	124,308	41
School Teacher	6,829	24	99,104	25	6,259	231	112,192	42
General Labourer	3	5386			110,428	12	110,431	43
Shoemaker	10,306	20	55,169	45	42,947	34	108,422	44
cultivateur	107,087	3					107,087	45
Grocer	2,501	74	53,303	46	50,289	28	106,093	46
Farm Labourer	16,080	14	5,850	232	84,136	15	106,066	47
No Occupation	2	6228	37,545	65	67,237	20	104,784	48
General Servant	22	1400	184	3496	104,538	13	104,744	49
Machinist	4,882	38	69,415	37	20,557	77	94,854	50
Seamstress	6,017	26	65,783	40	21,042	72	92,842	51
Teamster	3,498	53	88,603	28	32	19567	92,133	52
Farm Labor			85,567	30	262	3421	85,829	53
Wife	1	28103	5,817	233	77,301	17	83,119	54
Attends School	1,078	123	80,817	32	189	4536	82,084	55
Charwoman	26	1253	86	6267	81,394	16	81,506	56
Student	25,090	10	50,735	51	5,024	281	80,849	57
Works In Cotton Mill			80,420	33	37	17323	80,457	58

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
Work On Farm			79,818	34	244	3621	80,062	59
Merchant	11,924	18	62,584	42	4,208	353	78,716	60
Teacher	5,222	35	67,264	39	5,019	282	77,505	61
Stone Mason	1,281	113	41,467	57	33,580	46	76,328	62
Going To School	44,235	6	31,486	75	4	125417	75,725	63
Tailoress	5,084	36	29,144	81	41,190	36	75,418	64
Bricklayer	1	21763	4,386	282	70,739	19	75,126	65
Gardener	2,811	65	19,104	106	51,678	26	73,593	66
None	15	1737	32,740	70	39,783	40	72,538	67
Baker	2,738	68	32,609	71	35,920	45	71,267	68
Printer	3,487	55	52,978	47	14,658	110	71,123	69
Boarder	98	545	67,638	38	1,104	1072	68,840	70
Milliner	2,990	61	29,407	80	35,932	44	68,329	71
Joiner	1,327	109	2,345	483	64,206	22	67,878	72
Fisherman	18,077	13	25,626	87	23,069	67	66,772	73
Retired Farmer	325	271	44,582	53	20,374	79	65,281	74
Farmers Son	8	2816	19	19399	63,701	23	63,728	75
Cooper	3,815	46	43,935	54	13,441	121	61,191	76
Nurse	632	174	34,437	69	24,586	62	59,655	77
Physician	1,377	103	57,329	43	526	1937	59,232	78
Shoe Maker	9	2565	38,327	63	20,766	75	59,102	79
Domestic	679	165	34,728	67	21,392	70	56,799	80
Cabinet Maker	2,728	69	27,458	84	25,180	58	55,366	81
House Work			52,737	48	2,444	543	55,181	82
Lawyer	1,326	110	51,434	50	436	2260	53,196	83
Sailor	2,772	67	40,733	59	7,688	193	51,193	84
Farmers Wife			269	2618	50,857	27	51,126	85
Day Laborer			50,462	52	97	7877	50,559	86
Engineer	3,495	54	43,927	55	2,990	464	50,412	87
Housemaid	116	488	1,096	896	48,010	30	49,222	88
Miller	3,936	42	38,453	62	5,153	275	47,542	89

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
General Laborer			1,109	890	45,522	32	46,631	90
General Serv					46,215	31	46,215	91
Errand Boy	61	725	5,442	242	40,648	39	46,151	92
Cigar Maker	743	151	38,105	64	6,492	224	45,340	93
Plasterer	1,607	95	21,817	98	21,845	69	45,269	94
Barber	1,197	117	40,944	58	1,397	871	43,538	95
Working On Farm			42,882	56	97	7886	42,979	96
Labor	1	12792	40,593	60	243	3630	40,837	97
Agricultural Labourer					40,814	38	40,814	98
Carter	1,507	99	1,601	650	37,235	43	40,343	99
Domestic Serv	1,743	90	315	2342	37,551	42	39,609	100
K. HOUSE			38,661	61			38,661	101
Mason	3,233	58	7,074	197	25,599	54	35,906	102
Works In Woolen Mill			35,680	66	4	116678	35,684	103
journalier	35,430	8					35,430	104
WORKS IN FARM			34,678	68			34,678	105
Housework			30,814	77	3,862	384	34,676	106
Plumber	742	152	12,780	131	20,290	80	33,812	107
Book Keeper	3,365	56	25,422	88	4,830	298	33,617	108
Porter	562	188	13,281	128	18,584	90	32,427	109
Farmer Son	21,313	11	28	14463	11,035	139	32,376	110
Laborer On Farm			32,074	72	131	6156	32,205	112
Ag Labourer					32,075	47	32,075	113
Goes To School	10	2281	31,863	73	29	21366	31,902	114
Brick Mason	88	589	31,788	74	13	42605	31,889	115
Carman	2	8964	2,086	532	29,751	48	31,839	116
Coachman	65	703	13,523	125	18,187	95	31,775	117
Teaching School			31,436	76	2	265306	31,438	118
Farm Work			29,689	78	385	2512	30,074	119
Saloon Keeper	260	309	29,510	79	3	174561	29,773	120
Harness Maker	1,981	87	23,295	92	4,083	366	29,359	121

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
Bookkeeper			27,376	85	1,552	800	28,928	122
Housewife	42	897	8,545	176	20,035	82	28,622	123
Boarding	15	1786	28,474	82	7	68027	28,496	124
Seaman	5,926	27	11,272	140	10,482	146	27,680	125
Farms			27,166	86	2	187678	27,168	126
Farm Servant (Indoor)					27,160	50	27,160	127
Wheelwright	640	172	7,007	199	19,114	87	26,761	128
Iron Moulder	125	471	7,860	185	18,587	89	26,572	129
Mariner	1,993	86	3,549	332	20,957	73	26,499	130
House Carpenter	322	276	21,738	99	4,106	362	26,166	131
Farmers Daur					25,842	52	25,842	132
General Servant Domestic					25,720	53	25,720	133
Moulder	2,096	82	21,931	97	1,683	747	25,710	134
Cotton Winder			17	20820	25,595	55	25,612	135
Worsted Weaver			23	16884	25,452	56	25,475	136
Pupil Teacher			2	145718	25,288	57	25,290	137
Gen Lab	20	1469	19	19771	24,903	59	24,942	139
Railway Porter	15	1743	1	263586	24,820	60	24,836	140
Retail Grocer	1	21674	24,689	90	49	13865	24,739	141
Field Hand			24,566	91	97	7873	24,663	142
Hawker	4	4626	102	5516	24,527	63	24,633	143
Druggist	1,369	104	22,443	95	609	1712	24,421	144
Stone Cutter	1,308	112	20,951	100	1,832	708	24,091	145
Lab	6,292	25	3,194	362	14,330	114	23,816	146
In School	384	238	23,234	93	172	4912	23,790	147
Cotton Spinner	38	965	348	2150	23,388	66	23,774	148
Engine Fitter	24	1308	6	45750	23,670	64	23,700	149
Minister	1,441	101	22,067	96	104	7474	23,612	151
Drapers Assistant					23,506	65	23,506	152
Huckster	19	1537	22,830	94	317	2972	23,166	154
Son	15,557	16	1,261	799	6,141	239	22,959	156

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
Farm Servant	1,328	108	717	1228	20,875	74	22,920	157
Wagon Maker	1,454	100	20,825	101	148	5559	22,427	158
Infant			361	2091	21,908	68	22,269	159
Boot Maker	91	575	3,177	363	18,987	88	22,255	160
Silk Weaver	2	8801	2,933	393	18,465	91	21,400	163
Servant Domestic			590	1428	20,740	76	21,330	165
Cotton Operative			34	12573	21,136	71	21,170	166
Weaver	2,700	71	10,885	145	6,893	209	20,478	169
Woollen Weaver			9	35369	20,424	78	20,433	170
Warehouseman	14	1820	217	3075	20,088	81	20,319	171
Draper	27	1226	25	15778	19,943	83	19,995	174
Agricultural Laborer			26	15501	19,834	84	19,860	175
Shepherd			2,572	442	17,208	99	19,780	176
Grocers Assistant					19,730	85	19,730	177
Lodging House Keeper	1	27525	75	6959	19,170	86	19,246	180
Farmers Daughter			19	19670	18,437	92	18,456	183
Police Constable	16	1676	26	15248	18,259	94	18,301	185
General Lab					18,265	93	18,265	186
Ship Carpenter	2,366	77	9,845	159	5,926	246	18,137	187
Dom Serv			72	7199	17,794	97	17,866	189
Bricklayers Labourer					17,799	96	17,799	190
Worsted Spinner			16	22569	16,959	100	16,975	192
Farmer's Son	15,719	15	23	17063	1,172	1029	16,914	193
Tanner	1,554	97	10,739	148	4,335	341	16,628	198
Tinsmith	3,209	59	12,496	132	857	1306	16,562	199
Gentleman	3,866	45	4,303	285	4,197	354	12,366	251
Serv	1,868	89	561	1488	8,634	181	11,063	267
Carriage Maker	2,004	85	7,862	184	236	3729	10,102	298
Brick Layer	1,634	92	6,048	224	1,996	652	9,678	311
Store Clerk	2,261	79	6,413	213	15	37100	8,689	340
Agent	2,340	78	3,093	372	2,556	527	7,989	356

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
Store Keeper	3,531	52	4,017	297	331	2869	7,879	360
Not Given	7,781	21	45	10275	15	37578	7,841	362
Fmr Son	7,483	22			1	499879	7,484	377
Factory Hand	3,143	60	1,357	750	2,411	550	6,911	399
cultivateur son	6,904	23					6,904	400
Labour	2,466	75	2,156	519	2,000	650	6,622	415
Domestique	5,614	31			4	117903	5,618	492
menuisier	5,030	37					5,030	534
Rentier	4,789	39	23	17036	28	21865	4,840	553
Hunter	3,589	51	1,094	900	2	190660	4,685	567
Journeyman	4,263	41	40	11094	57	12202	4,360	610
Hotelkeeper	3,877	44	229	2948	56	12302	4,162	630
commis	3,894	43					3,894	660
cordonnier	3,800	47					3,800	674
couturiŠre	3,799	48					3,799	675
Servante	3,760	49	1	238301	1	485394	3,762	681
Trader	1,630	94	2,043	545	8	65078	3,681	693
Voyageur	3,675	50			1	830923	3,676	694
Lumber Man	2,951	62	666	1311	1	345955	3,618	703
forgeron	3,330	57					3,330	751
MARCHAND	2,869	63	2	110331			2,871	844
FMR LAB	2,843	64					2,843	851
navigateur	2,792	66					2,792	857
fermier	2,709	70					2,709	882
cultivateur fils	2,597	72					2,597	919
Institutrice	2,520	73	2	146061	2	318709	2,524	944
Cobbler	1,632	93	786	1158	47	14312	2,465	966
charretier	2,401	76					2,401	985
S.	2,101	80	31	13648	14	40042	2,146	1067
Religieuse	2,097	81			10	51155	2,107	1088
CHARPENTIER	2,057	83	1	212114			2,058	1117

Occupation	Canada		United States		Great Britain		All	
	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank
pˆcheur	2,010	84					2,010	1144
cultivateur's son	1,901	88					1,901	1201
commer‡ant	1,667	91					1,667	1322
Etudiant	1,564	96			1	1195158	1,565	1397
BLACK SHOP	1,522	98	1	528589			1,523	1433

Selected references

Carter, Susan B., Roger L. Ransom, and Richard Sutch. "The Historical Labor Statistics Project at the University of California." *Historical Methods* 24, no. 1 (1991): 52-65.

Cohen, Patricia Cline. *A calculating people: the spread of numeracy in early America*. Chicago: University of Chicago Press, 1982.

Foreman-Peck, James. *New perspectives on the late Victorian economy: essays in quantitative economic history, 1860-1914*. Cambridge: Cambridge University Press, 1991.

Friedman, Walter A. *Birth of a salesman: the transformation of selling in America*. Cambridge, MA: Harvard University Press, 2004.

Magnuson, Diana, and Miriam King. "Enumeration Procedures." *Historical Methods* 28, no. 1 (1995): 27-32.

Sobek, Matthew. "Work, Status, and Income: Men in the American Occupational Structure since the Late Nineteenth Century." *Social Science History* 20, no. 2 (1996): 169-207.

van Leeuwen, Marco H.D., Ineke Maas, and Andrew Miles. "Creating a Historical International Standard Classification of Occupations." *Historical Methods* 37, no. 4 (2004): 186-197.

———. *Historical International Standard Classification of Occupations*. Leuven: Leuven University Press, 2002.

Vikström, Par, Sorën Edvinsson, and Anders Brändström. "Longitudinal Databases — Sources For Analyzing The Life-Course: Characteristics, Difficulties And Possibilities." *History and Computing* 14, no. 1/2 (2006): 109-128.